ERDC/CHL CHETN-I-91
July 2016

US Army Corps of Engineers®

Evaluation Statistics Computed for the Wave Information Studies (WIS)

by Mary A. Bryant, Tyler J. Hesser, and Robert E. Jensen

PURPOSE: This Coastal and Hydraulics Engineering Technical Note (CHETN) describes the statistical metrics used by the Wave Information Studies (WIS) and produced as part of the model evaluation process.

INTRODUCTION: The objective of the WIS is to provide coastal wave hindcast model estimations, wave analyses products, and decision support tools nationwide. These wave estimates are hindcast using high-quality climatology data and third-generation wave models (i.e., WAM, Komen et al. 1994; WAVEWATCH III, Tolman 2014). The resulting wave estimates (height, period, and direction) and directional spectral estimates are provided for a set of preselected, virtual gauge locations along the Pacific, Great Lakes, Gulf of Mexico, Atlantic, and Western Alaska coasts.

Estimates of wave climatology produced by ocean wave models, including those of WIS, are influenced by meteorological forcing parameters, representation of the geographic area (e.g., bathymetry), and inherent model physics and assumptions. An integral part of assessing the performance of these wave models is a quantitative evaluation comparing model estimates to wave measurements. As part of the WIS effort, this evaluation extends over a large spatial area, wave climate regimes, and meteorological events. One component of the evaluation process is the computation of summary statistics. These error statistics, or statistical metrics, include bias, root-mean-square error (RMSE), scatter index, symmetric slope, and correlation coefficient. Some of the earliest applications of these statistical metrics to wave model evaluation are found in Zambresky (1989) and Cardone et al. (1996). These statistics were calculated in the transition of the WAVEWATCH III model to the Naval Oceanographic Office (Rogers et al. 2012) and more recently to evaluate the National Centers for Environmental Prediction's operational wave forecasting system for Hurricane Sandy (Alves et al. 2015). With the ongoing development and widespread application of ocean wave models, methods to evaluate their performance comprehensively are needed. However, as discussed later, these evaluations are complicated by model studies defining slightly different statistical metrics yet referring to these metrics with the same name.

STATISTICS: In this section, definitions and interpretations of the various statistics computed by WIS are given. Within WIS, these statistics are computed for wind speed and the scalar statistical wave descriptors of wave height and both mean and peak period. Variations of these statistics are also computed for directional data, wind direction, and mean wave direction. For each of the definitions given below, X represents the observed measurements, and Y represents the corresponding modeled hindcast values in a series of N measurements.

Mean of measured (X) and hindcast (Y) parameters: $\bar{X} = \frac{1}{N}\sum X_i$ $\bar{Y} = \frac{1}{N}\sum Y_i$

Bias (hindcast-measured): $b = \frac{1}{N}\sum\left(Y_i - X_i\right)$

The bias is a representation of the model's mean, long-term error, where its value either indicates an average overestimation (positive) or underestimation (negative) compared to the measurements.

RMSE (demeaned): $RMSE = \sqrt{\frac{1}{N-1}\sum\left(Y_i - X_i - b\right)^2}$

The RMSE is a measure of the residuals between the model predictions and measured observations, where larger numbers indicate greater variance. Whereas the WIS definition of RMSE is corrected for the bias (demeaned), resulting in its equivalence to the standard deviation of the difference, other reports of RMSE include components of variance and bias (E_{RMS}) and may be normalized (NRMSE) (Ardhuin et al. 2010):

$$E_{RMS} = \sqrt{\frac{1}{N}\sum\left(Y_i - X_i\right)^2}$$

$$NRMSE = \sqrt{\frac{\sum\left(Y_i - X_i\right)^2}{\sum X_i^2}}$$

By presenting the RMSE as unbiased, a more complete picture of the error distribution is provided (Chai and Draxler 2014). However, one must use caution when comparing across different model applications as each study may compute a different definition for RMSE.

Scatter index (SI): $SI = \frac{RMSE}{\bar{X}}$

The SI is a normalized measure of error, often reported as a percent. Lower values of the SI are an indication of better model performance. Like the RMSE, ambiguities exist in the definition of the scatter index, with authors either defining it as the standard deviation of the errors (i.e., demeaned RMSE) divided by the mean of the observations (Mentaschi et al. 2013), as done by WIS, or defining it as the E_{RMS} (defined above) divided by the mean of the observations (Ris et al. 1999; Rogers et al. 2012; Akpinar et al. 2012).

Symmetric slope: $sym\,r = \sqrt{\frac{\sum Y_i^2}{\sum X_i^2}}$

The symmetric slope, *sym r*, is the coefficient of linear regression constrained to pass through the origin (*y*-intercept = 0) and is ideally close to 1.0. Slopes greater than 1.0 indicate a consistent overestimation, and slopes under 1.0 indicate a consistent underestimation by the model.

Correlation coefficient: $corr = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}}$

The Pearson correlation coefficient, *corr*, is a measure of the degree of linear dependence between the model and the observations (Rogers et al. 2012). A perfect positive linear relationship (i.e., as the value of one variable increases, the value of the other variable increases) has a value of 1.0 while no linear relationship is indicated by a value of 0.0. The correlation coefficient can also measure the degree of decreasing linear relationship (-1.0 indicates a perfect negative linear relationship); however, decreasing linear relationships are not applicable to the WIS evaluation.

SENSITIVITY: To illustrate the sensitivity of these statistical metrics, a series of sensitivity tests to known modifications of a base dataset are shown in Figure 1. To perform these sensitivity tests, the observation dataset was duplicated and then modified by a defined additive amount and/or time shift to generate artificial *model* results. This process allowed the observation signal to be maintained while exploring the response of the statistics to common model errors. These sensitivity tests were performed for National Data Buoy Center (NDBC) buoy 44065 located in the New York Harbor Entrance. The time series plots, shown in the left column of Figure 1, compare time series of the *modeled* (black line) and observed (red dots) zero-moment wave heights (H_{mo}). The plots on the right in Figure 1 compare the time-paired modeled H_{mo} on the vertical axis to the measured H_{mo} on the horizontal axis. Within these plots, the diagonal black line is the *best fit line*, and the closer a dot lies to the best fit line, the better the model hindcast is of that measurement. The distance of the dots above and below the black line indicates the degree of overestimation or underestimation by the model, respectively. The blue line represents the symmetric regression line, given by the formula $Y = (sym\ r)X$. Moving from top to bottom within Figure 1, the panels are the following: the top panel (a) is a perfect model result (i.e., model identical to observations), the second panel (b) is a positive shift in the model by a constant value (0.3 meters [m]), the third panel (c) is a phase lag of the model (2 hours [hr]), the fourth panel (d) is a larger phase lag of the model (12 hr), and the bottom panel (e) is a combination of a bias (0.3 m) and a phase lag (2 hr).

The statistical results of these sensitivity tests are provided in Table 1. As expected, the perfect model has a bias, RMSE, and SI of 0.0 and a symmetric regression and correlation coefficient of 1.0. A constant additive of the model mean compared to the measurements results in a bias equal to the magnitude of the shift (0.3 m) and an increase in the symmetric regression (1.17). These statistics both indicate an overestimation of the model compared to the measurements. Note that both the RMSE and SI remain 0.0 because both statistics are demeaned. Although the actual values of the model and measurements differ, the correlation coefficient remains 1.0 as expected because the linear response between the model and measurements is identical. Lagging the model by 2 hr had little effect on the bias, symmetric regression, and correlation. Increasing the lag to 12 hr significantly lowered the correlation (0.703) because the linear relationship between the model and measurements weakened compared to the 2 hr lag, as shown by the increase in scatter along the line of best fit. However, the bias and symmetric regression remained approximately 0.0 and 1.0, respectively. The bias and symmetric regression are based only on the composition of the population sets. Since the data at the beginning and end of the time window were varying slowly, the bias and symmetric regressions were only changed slightly. Increasing the time lag beyond 12

hr such that larger events were purposely excluded from the model population was shown to slightly lower the symmetric regression. However, the regression remained above 0.9 as smaller wave heights dominated the population composition, which is common in wave modeling applications. Data with greater scatter along the line of best fit have an increase in the RMSE. This increase in the RMSE, and the corresponding increase in the SI, with model phasing is expected given that the RMSE is computed with data paired in time. The statistics for the bias and phase lag case are a superposition of their individual results.

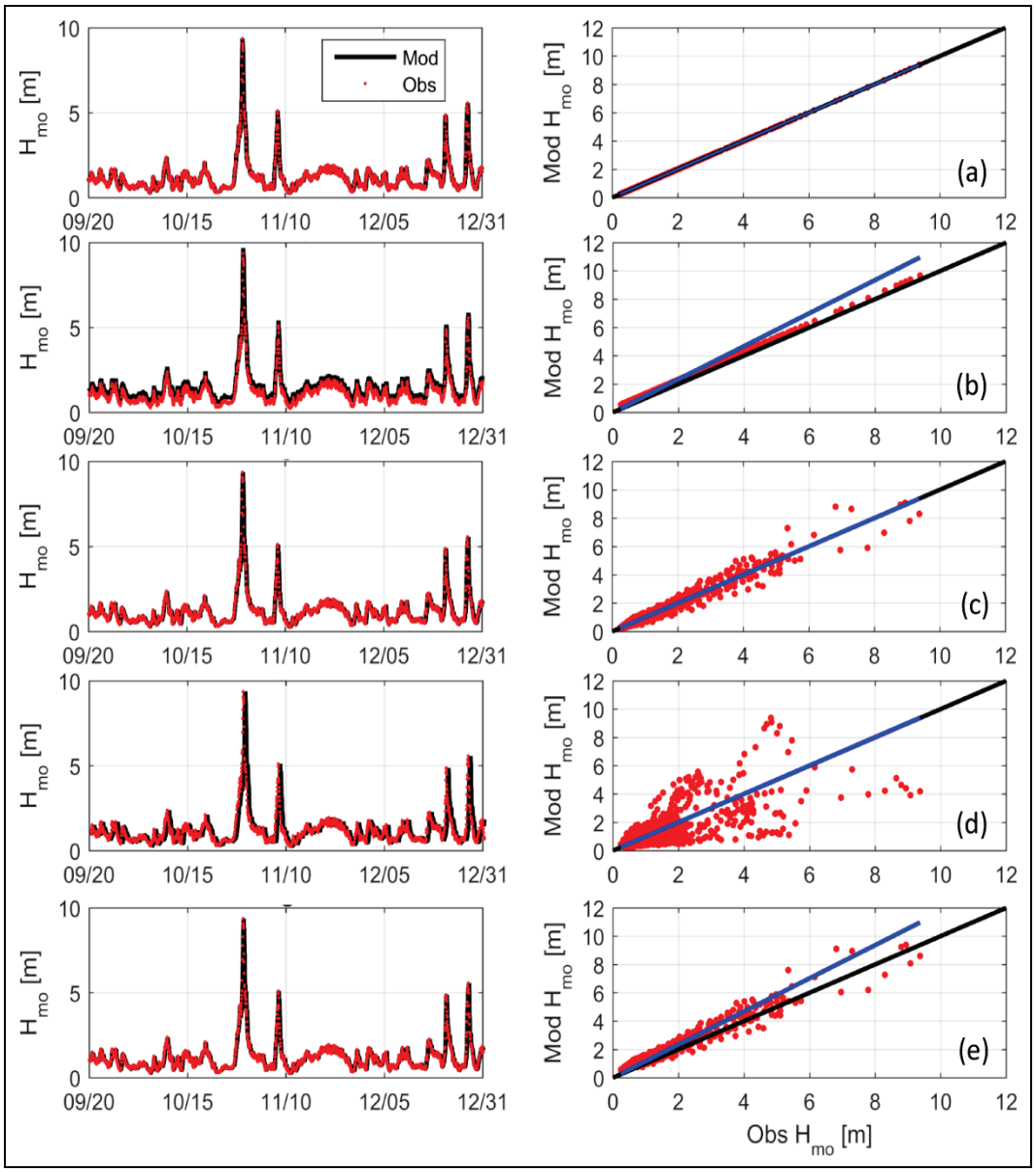

Figure 1. Statistics sensitivity to the following variations: (a) perfect model, (b) 0.30 m positive bias, (c) 2 hr phase lag, (d) 12 hr phase lag, and (e) combination of 0.30 m positive bias and 2 hr phase lag.

Table 1. Computed WIS statistics for various sensitivity tests.

Condition	Bias [m]	RMSE [m]	SI	Sym r	Corr
Perfect model	0.0	0.0	0.0	1.00	1.0
Bias (0.30 m)	0.30	0.0	0.0	1.17	1.0
Phase lag (2 hr)	0.0	0.19	16.05	1.00	0.976
Phase lag (12 hr)	0.0	0.68	57.25	1.00	0.703
Bias (0.30 m) and phase lag (2 hr)	0.30	0.19	16.05	1.17	0.976

In summary, a single statistic provides only limited information about a certain aspect of model performance. Adjusting the model results by a constant relative to the measurements was only reflected in the bias and the symmetric regression whereas a time shift in the model results noticeably affected the RMSE, SI, and correlation. Thus, considering a combination of metrics is required to broaden the error interpretation associated with model performance (Chai and Draxler 2014).

EXAMPLE: Figure 2 shows an example of a WIS evaluation, with model results compared to observations at NDBC 45007 in southern Lake Michigan from 31 May 2014 to 5 December 2014. The bias is small, approximately 0.09 m. The RMSE is approximately 0.28 m with a scatter index of 42.73. The correlation coefficient is 0.946. The symmetric regression is 1.05, indicating a 5% consistent overestimation of the observations by the model. Looking at the scatter plot, the wave climate is dominated by wave heights of approximately 2 m or less. The model results overestimate wave heights less than 1.0 m with scatter from 1.0 to 3.0 m roughly distributed evenly above and below the best fit line. Wave heights are slightly underestimated in the 3 to 4.8 m range. The model overestimates waves larger than 4.8 m by 0.5 m or more. The maximum wave height of the timeframe, approximately 6.6 m, is underestimated by WIS by approximately 1 m, as seen in the inset. Note that while the maximum modeled wave height is much closer in value to the observed peak wave, the model is slightly out of phase and thus results in a larger error.

LIMITATIONS: Condensing a set of error values to a single number will inevitably have limitations. For example, a net zero bias can occur when an overestimation of a large population of low wave conditions occurs in conjunction with an underestimation of a small population of larger storm conditions. This scenario can also result in symmetric regression values very close to 1.0 (Figure 3). The correlation coefficient is unable to discern differences in proportionality and/or constant additive differences between two variables, as demonstrated above (Willmott 1981). Both Mentaschi et al. (2013) and Willmott et al. (2009) suggest the sums-of-squares based errors, such as the RMSE and its variants, can be misleading and may not always be reliable to assess the accuracy of numerical models. The sensitivity of the RMSE to outliers is a common concern, especially when the outliers are not well represented in a smaller sample size. For instances of small mean wave heights, as in coastal applications, model errors can often approach the magnitude of the observations, elevating the scatter index in low wave conditions. For example, an RMSE of 0.2 m in H_{mo} seems reasonable, but if the mean measured H_{mo} is only 0.4 m, the scatter index attains a rather high value of 50%.

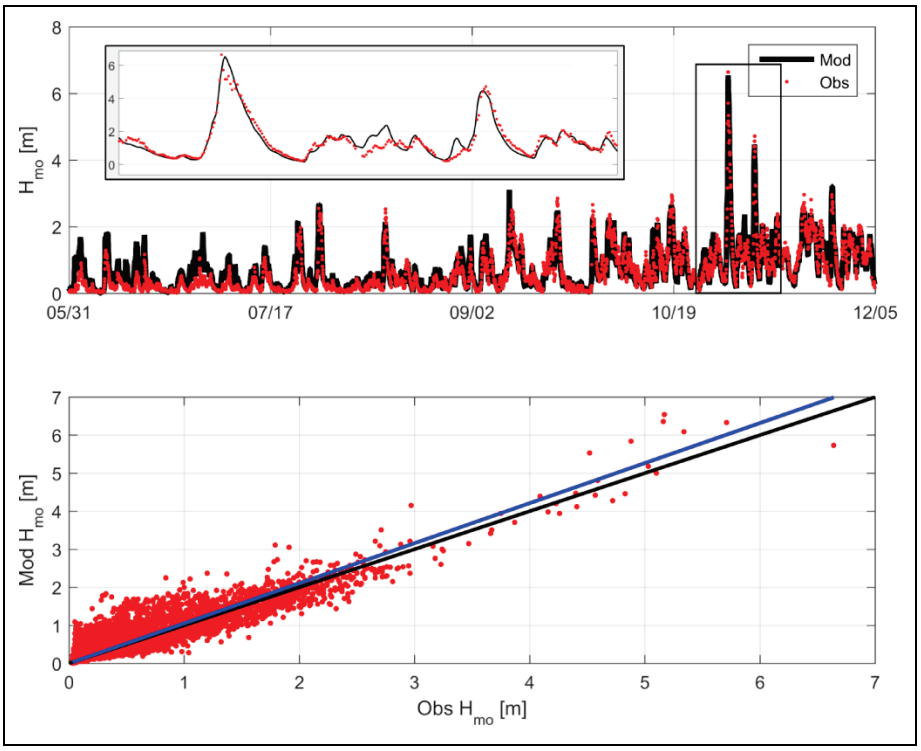

Figure 2. Evaluation of WIS results to NDBC Buoy 45007. Inset is enlarged maximum wave heights.

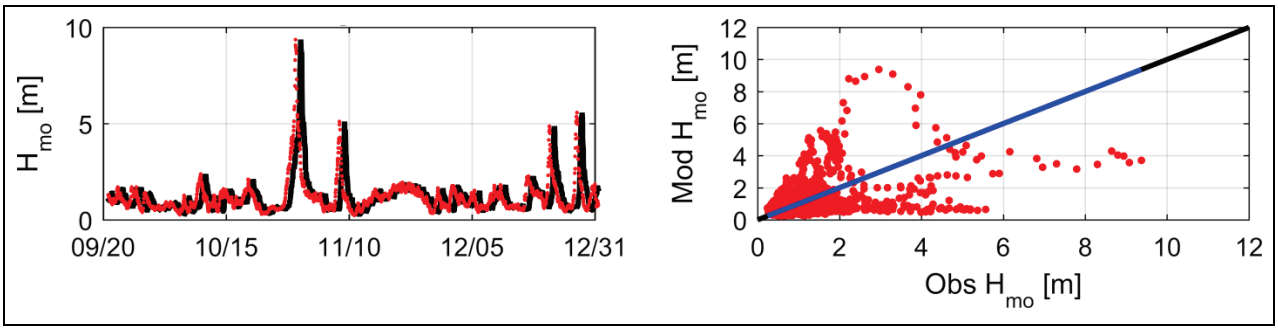

Figure 3. Example of scattered data (phase lag of 24 hr) that yield a bias and symmetric regression of approximately 0.0 and 1.0, respectively.

PERFORMANCE SCORES: Because the individual statistics may sometimes be misleading in assessing model performance, the investigation of concise overall performance scores as additional indicators is ongoing. One such performance score is the Willmott et al. (1985) index, defined as the following:

$$d_1 = 1 - \frac{\sum |Y_i - X_i|}{\sum \left(|Y_i - \bar{X}| + |X_i - \bar{X}| \right)}$$

This version of the Willmott index is based on the absolute values of the errors and is less sensitive to errors concentrated in outliers compared to its original formulation (Willmott 1981). Another

overall performance score considered is one computed by the Interactive Model Evaluation and Diagnostic System (IMEDS, Hanson et al. 2009). It normalizes the E_{RMS} and bias by the root-mean-square of the measurements and then averages these normalized error estimates:

$$X_{rms} = \sqrt{\frac{1}{N}\sum X_i^2}$$

$$p_{rms} = 1 - \frac{E_{RMS}}{X_{rms}}$$

$$p_{bias} = 1 - \frac{|b|}{X_{rms}}$$

$$P_s = \frac{p_{rms} + p_{bias}}{2}$$

An upper bound of 1.0 for both skill scores indicates perfect model performance. Considering again the sensitivity study shown in Figure 1, Table 2 is updated to show the Willmott and IMEDS skill scores for each case. Whereas the other statistics demonstrate the general behavior of sensitivity to either bias or time, the performance scores respond to changes in both. Additionally, the responses of the performance scores when bias and phase lag are in combination are completely distinct from their responses to the conditions individually, unlike the other statistics.

Table 2. Computed WIS statistics and performance scores for various sensitivity tests.

Condition	Bias [m]	RMSE [m]	SI	Sym r	Corr	Willmott et al. (1985)	IMEDS
Perfect model	0.0	0.0	0.0	1.00	1.0	1.0	1.0
Bias (0.30 m)	0.30	0.0	0.0	1.17	1.0	0.71	0.80
Phase lag (2 hr)	0.0	0.19	16.05	1.00	0.976	0.90	0.94
Phase lag (12 hr)	0.0	0.68	57.25	1.00	0.703	0.65	0.77
Bias (0.30 m) and phase lag (2 hr)	0.30	0.19	16.05	1.17	0.976	0.69	0.78

Figure 4 further investigates the behavior of these performance indices relative to the other statistical metrics. These four panels show the evaluation of the statistics for increasing bias and phase lag for two buoys, NDBC 44065 (top two panels) and 45007 (bottom two panels). The bias in Figure 4 is represented using the normalized bias (NBias), which is the bias (b) divided by the mean of the observations ($\overline{X}$). Again, as the bias increases, only the symmetric regression and the Willmott and IMEDS performance indices change, as shown in the first and third panel for the two buoys. However, the response of the indices is considerably different, with IMEDS declining linearly and Willmott declining exponentially. IMEDS initially produces slightly higher skill values than Willmott until approximately 0.8 NBias; thereafter, IMEDS continues to

decline to a negative score. The Willmott index's lower limit is 0.0, which it approaches much more slowly at high biases. For the phase lag, there is only a noticeable change in the scatter index and correlation coefficient, as shown in the second and fourth panel. The response of the Willmott and IMEDS is similar although the IMEDS index appears to be more lenient.

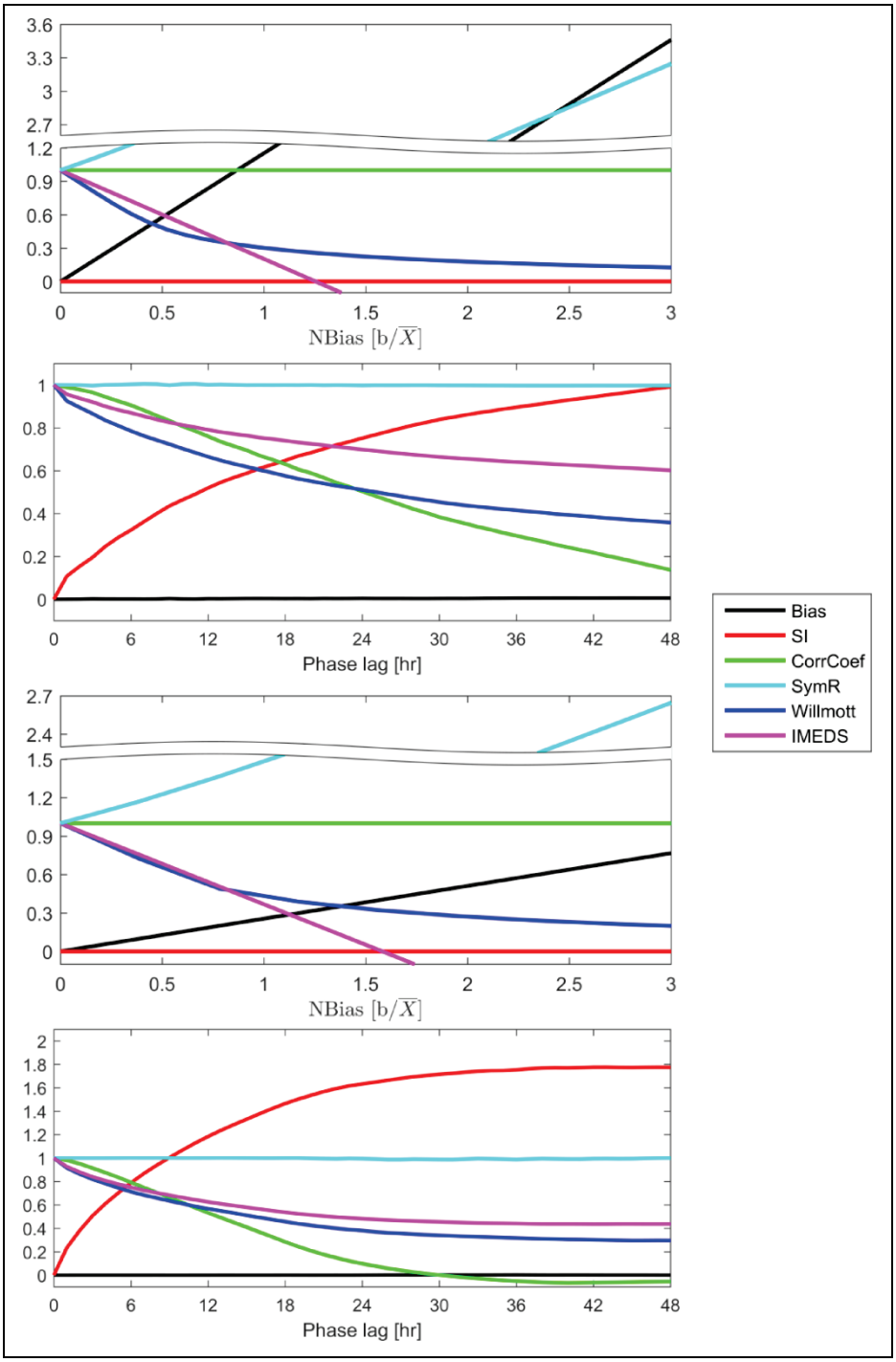

Figure 4. Response of WIS statistical metrics and considered performance scores to progressive changes in bias and phase for two buoys, NDBC 44065 (top two panels) and 45007 (bottom two panels).

CONCLUSIONS: In this technical note, an overview of the statistical metrics computed by the Wave Information Studies is given. These statistical metrics provide a comprehensive evaluation of hindcast performance. The statistics computed by WIS include bias, demeaned RMSE, scatter index, symmetric slope, and correlation coefficient. Sensitivity studies revealed that the bias and symmetric slope are sensitive to changes in the mean of the model results compared to the measurements. The RMSE, SI, and correlation coefficient are sensitive to time shifts, lag or lead, in the model results relative to the measurements.

Including performance scores, such as Willmott et al. (1985) and IMEDS, complements the WIS evaluation as these metrics suggest an overall skill assessment and are sensitive to both bias and time shifts of the model with respect to measurements. The Willmott index has a lower limit of 0.0, which alters its behavior from that of IMEDS at high biases. The interpretation of these statistics with respect to model performance is subjective—for example, what performance score is indicative of *good* or *poor* model performance? One way to eliminate the subjective nature of the evaluation process is a comparative evaluation. However, comparisons across models are difficult because the statistical definitions are not always defined or are shown to vary, such as with the RMSE and SI. To overcome these challenges, the wave model community should make strides in standardizing statistical metrics to advance the objective evaluation of numerical models.

ADDITIONAL INFORMATION: This CHETN was prepared as part of Wave Information Studies (WIS) work unit in the Coastal Ocean Data System Program and was written by Mary A. Bryant (*Mary.Bryant@usace.army.mil*), Tyler J. Hesser (*Tyler.Hesser@usace.army.mil*), and Robert E. Jensen (*Robert.E.Jensen@usace.army.mil*) of the U.S. Army Engineer Research and Development Center (ERDC), Coastal and Hydraulics Laboratory (CHL). The Program Manager is Dr. Jeffrey P. Waters, and the Technical Directors are William Curtis and W. Jeffrey Lillycrop. This CHETN should be cited as follows:

> Bryant, M. A., T. J. Hesser, and R. E. Jensen. 2016. *Evaluation statistics computed for the Wave Information Studies (WIS).* ERDC/CHL CHETN-I-91. Vicksburg, MS: U.S. Army Engineer Research and Development Center.

REFERENCES

Akpinar, A., G. Ph. van Vledder, M. İ. Kömürcü, and M. Özger. 2012. Evaluation of the numerical wave model (SWAN) for wave simulation in the Black Sea. *Continental Shelf Research* 50–51: 80–99.

Alves, J.-H. G. M., S. Stripling, A. Chawla, H. Tolman, and A. van der Westhuysen. 2015. Operational wave guidance at the U.S. National Weather Service during Tropical/Post-Tropical Storm Sandy, October 2012. *Monthly Weather Review* 143:1687–1702.

Ardhuin, F., E. Rodger, A. V. Babanin, J. F. Filipot, R. Magne, A. Roland, A. van der Westhuysen, P. Queffeulou, J. M. Lefevre, L. Aouf, and F. Collard. 2010. Semiempirical dissipation source functions for ocean waves. Part I: Definition, calibration, and validation. *Journal of Physical Oceanography* 40:1917–1941.

Cardone, V. J., R. E. Jensen, D. T. Resio, V. R. Swail, and A. T. Cox. 1996. Evaluation of contemporary ocean wave models in rare extreme events: The "Halloween Storm" of October 1991 and the "Storm of the Century" of March 1993. *Journal of Atmospheric and Oceanic Technology* 13:198–230.

Chai, T., and R. R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7:1247–1250.

Hanson, J. L., B. A. Tracy, H. L. Tolman, and R. D. Scott. 2009. Pacific hindcast performance of three numerical wave models. *Journal of Atmospheric and Oceanic Technology* 26:1614–1633.

Komen, G. J., L. Cavaleri, M. Donelan, K. Hasselmann, S. Hasselmann, P. A. E. M. Janssen. 1994. *Dynamics and modelling of ocean waves*. New York: Cambridge Univeristy Press.

Mentaschi, L., G. Besio, F. Cassola, and A. Mazzino. 2013. Problems in RMSE-based wave model validations. *Ocean Modelling* 72:53–58.

Ris, R. C., L. H. Holthuijsen, and N. Booij. 1999. A third-generation wave model for coastal regions, 2. Verification. *Journal of Geophysical Research: Oceans* 104 (C4):7667–7681.

Rogers, W. E., J. D. Dykes, and D. Wang. 2012. *Validation test report for WAVEWATCH III*. NRL/MR/7320--12-9425. Stennis Space Center, Mississippi: Naval Research Laboratory.

Tolman, H. L. 2014. User manual and system documentation of WAVEWATCH III version 4.18. NOAA/NWS/NCEP/MMAB Technical Note 316. http://polar.ncep.noaa.gov/waves/wavewatch/manual.v4.18.pdf

Willmott, C. J. 1981. On the validation of models. *Physical Geography* 2(2):184–194.

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe. 1985. Statistics for evaluation and comparison of models. *Journal of Geophysical Research: Oceans* 90 (C5):8995–9005.

Willmott, C. J., K. Matsuura, and S. M. Robeson. 2009. Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment* 43:749–752.

Zambresky, L. 1989. *A verification study of the global WAM model, December 1987–November 1988*. Technical Report No. 63. Reading, England: European Centre for Medium-Range Weather Forecasts.